Package ‘DWDLargeR’

February 6, 2018

Version 0.1-0

Date 2018-02-06

Title Fast Algorithms for Large Scale Generalized Distance Weighted Discrimination

Author Xin-Yee Lam, J.S. Marron, Defeng Sun, and Kim-Chuan Toh

Maintainer Kim-Chuan Toh <mattohkc@nus.edu.sg>

Depends R (>= 2.10), Matrix, SparseM

Imports methods, stats

Description Solving large scale distance weighted discrimination.
The main algorithm is a symmetric Gauss-Seidel based alternating direction method of multipliers (ADMM) method. See Lam, X.Y., Marron, J.S., Sun, D.F., and Toh, K.C. (2018) <arXiv:1604.05473> for more details.

License GPL-2

URL https://arxiv.org/pdf/1604.05473.pdf

NeedsCompilation no

Repository CRAN

Date/Publication 2018-02-06 10:48:04 UTC

R topics documented:

DWDLargeR-package *Fast Algorithms for Large Scale Generalized Distance Weighted Discrimination*

Description

Solving large scale distance weighted discrimination. The main algorithm is a symmetric Gauss-Seidel based alternating direction method of multipliers (ADMM) method.

Details

The package DWDLargeR contains two main functions:
penaltyParameter
genDWD

Author(s)

Xin-Yee Lam, J.S. Marron, Defeng Sun, and Kim-Chuan Toh

References

Lam, X.Y., Marron, J.S., Sun, D.F., and Toh, K.C. (2018) “Fast algorithms for large scale generalized distance weighted discrimination", *Journal of Computational and Graphical Statistics*, forthcoming.
https://arxiv.org/abs/1604.05473

genDWD *Solve the generalized distance weighted discrimination (DWD) model.*

Description

Solve the generalized DWD model by using a symmetric Gauss-Seidel based alternating direction method of multipliers (ADMM) method.

Usage

```
genDWD(X,y,C,expon, tol = 1e-5, maxIter = 2000, method = 1, printDetails = 0,
              rmzeroFea = 1, scaleFea = 1)
```

Arguments

`X` A d x n matrix of n training samples with d features.

`y` A vector of length n of training labels. The element of `y` is either -1 or 1.

`C` A number representing the penalty parameter for the generalized DWD model.

`expon` A positive number representing the exponent q of the residual r_i in the generalized DWD model. Common choices are `expon = 1,2,4`.

`tol` The stopping tolerance for the algorithm. (Default = 1e-5)

`maxIter` Maximum iteration allowed for the algorithm. (Default = 2000)

`method` Method for solving generalized DWD model. The default is set to be 1 for the highly efficient sGS-ADMM algorithm. User can also select `method = 2` for the directly extended ADMM solver.

`printDetails` Switch for printing details of the algorithm. Default is set to be 0 (not printing).

`rmzeroFea` Switch for removing zero features in the data matrix. Default is set to be 1 (removing zero features).

`scaleFea` Switch for scaling features in the data matrix. This is to make the features having roughly similar magnitude. Default is set to be 1 (scaling features).

Details

This is a symmetric Gauss-Seidel based alternating method of multipliers (sGS-ADMM) algorithm for solving the generalized DWD model of the following formulation:

$$\min \sum_i \theta_q(r_i) + Ce^T x_i$$

subject to the constraints

$$Z^T w + \beta y + \xi - r = 0, ||w|| <= 1, \xi >= 0,$$

where $Z = Xdiag(y)$, e is a given positive vector such that $||e||_\infty = 1$, and θ_q is a function defined by $\theta_q(t) = 1/t^q$ if $t > 0$ and $\theta_q(t) = \infty$ if $t <= 0$.

Value

A list consists of the result from the algorithm.

`w` The unit normal of hyperplane that distinguishes the two classes.

`beta` The distance of the hyperplane to the origin (β in the above formulation).

`xi` A slack variable of length n for the possibility that the two classes may not be separated cleanly by the hyperplane (ξ in the above formulation).

`r` The residual $r := Z^T w + \beta y + \xi$.

`alpha` Dual variable of the generalized DWD model.

`info` A list consists of the information from the algorithm.

`runhist` A list consists of the run history throughout the iterations.

Author(s)

Xin-Yee Lam, J.S. Marron, Defeng Sun, and Kim-Chuan Toh

References

Lam, X.Y., Marron, J.S., Sun, D.F., and Toh, K.C. (2018) “Fast algorithms for large scale generalized distance weighted discrimination", *Journal of Computational and Graphical Statistics*, forthcoming.
https://arxiv.org/abs/1604.05473

Examples

```
# load the data
data("mushrooms")
# calculate the best penalty parameter
C = penaltyParameter(mushrooms$X,mushrooms$y,expon=1)
# solve the generalized DWD model
result = genDWD(mushrooms$X,mushrooms$y,C,expon=1)
```

mushrooms *Classification data from Audobon Society Field Guide (1981).*

Description

This data set includes descriptions of hypothetical samples corresponding to 23 species of gilled mushrooms in the Agaricus and Lepiota Family (pp. 500-525). Each species is identified as definitely edible, definitely poisonous, or of unknown edibility and not recommended. This latter class was combined with the poisonous one. The Guide clearly states that there is no simple rule for determining the edibility of a mushroom; no rule like “leaflets three, let it be” for Poisonous Oak and Ivy.

Usage

```
data(mushrooms)
```

Format

List containing a 112x8124 matrix of 8124 training samples with 112 features; and a vector of length 8124 training labels.

Source

The data could be downloaded from the UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/Mushroom

References

Lichman, M. (2013). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science.

`penaltyParameter` *Compute the penalty parameter for the model.*

Description

Find the best penalty parameter C for the generalized distance weighted discrimination (DWD) model.

Usage

```
penaltyParameter(X,y,expon,rmzeroFea = 1, scaleFea = 1)
```

Arguments

`X`	A d x n matrix of n training samples with d features.
`y`	A vector of length n of training labels. The element of `y` is either -1 or 1.
`expon`	A positive number representing the exponent q of the residual r_i in the generalized DWD model. Common choices are `expon = 1,2,4`.
`rmzeroFea`	Switch for removing zero features in the data matrix. Default is set to be 1 (removing zero features).
`scaleFea`	Switch for scaling features in the data matrix. This is to make the features having roughly similar magnitude. Default is set to be 1 (scaling features).

Details

The best parameter is empirically found to be inversely proportional to the typical distance between different samples raised to the power of $(expon + 1)$. It is also dependent on the sample size n and feature dimension d.

Value

A number which represents the best penalty parameter for the generalized DWD model.

Author(s)

Xin-Yee Lam, J.S. Marron, Defeng Sun, and Kim-Chuan Toh

References

Lam, X.Y., Marron, J.S., Sun, D.F., and Toh, K.C. (2018) “Fast algorithms for large scale generalized distance weighted discrimination", *Journal of Computational and Graphical Statistics*, forthcoming.
https://arxiv.org/abs/1604.05473

Examples

```
# load the data
data("mushrooms")
# calculate the best penalty parameter
C = penaltyParameter(mushrooms$X,mushrooms$y,expon=1)
```

Index